

# Queued cross-bar network models for replication and coded storage systems

Ulric J. Ferner, Neda Aboutorab, Parastoo Sadeghi, Muriel Médard

**Abstract**—Coding techniques may be useful for data center data survivability as well as for reducing traffic congestion. We present a queued cross-bar network (QCN) method that can be used for traffic analysis of both replication/uncoded and coded storage systems. We develop a framework for generating QCN rate regions (RRs) by analyzing their conflict graph stable set polytopes (SSPs). In doing so, we apply recent results from graph theory on the characterization of particular graph SSPs. We characterize the SSP of QCN conflict graphs under a variety of traffic patterns, allowing for their efficient RR computation. For uncoded systems, we show how to compute RRs and find rate optimal scheduling algorithms. For coded storage, we develop a RR upper bound, for which we provide an intuitive interpretation. We show that the coded storage RR upper bound is achievable in certain coded systems in which drives store sufficient coded information, as well in certain dynamic coding systems. Numerical illustrations show that coded storage can result in gains in RR volume of approximately 50%, averaged across traffic patterns.

## I. INTRODUCTION

The continued growth in data center (DC) demand worldwide is driving the development of new DC architectures and data management techniques. Two key parameters of data management in DCs are the survivability of data in the event of node failures and constant availability of data. Significant academic literature has focused on data survivability in the form of regenerating codes. See [1] and references therein. In enterprise-level DCs temporary drive unavailability dominates permanent failure by a factor of nine to one [2]. In this paper we concentrate on data unavailability due to traffic congestion.

We consider physical storage networks as illustrated in Fig. 1. Chunks are fixed-size file subsets. A network of drives, where each drive is either a hard disk drive (HDD), solid state drive (SSD), or RAM cache, stores some number of file chunks. Outside users send read requests for file chunks to the drive network, and drives process read requests and send chunks back to users. Time is slotted and in each timeslot the network is constrained as to how many users each drive can transmit a stored chunk to; we refer to these constraints as traffic patterns.

This material is based upon work supported by the Martin Family Society of Fellows for Sustainability at MIT, by BAE Systems National Security Solutions Inc., under award 739532-SLIN 0004, and the Australian Research Councils Discovery Projects funding scheme (project no. DP120100160). U.J. Ferner and M. Médard are currently with the Research Laboratory for Electronics, Massachusetts Institute of Technology, Room 36-512, 77 Massachusetts Avenue, Cambridge, MA 02139 (e-mail: {uferner, medard}@mit.edu). P. Sadeghi and N. Aboutorab are with the Research School of Information Sciences and Engineering, The Australian National University, Canberra, Australia (e-mail: {parastoo.sadeghi, neda.aboutorab}@anu.edu.au).

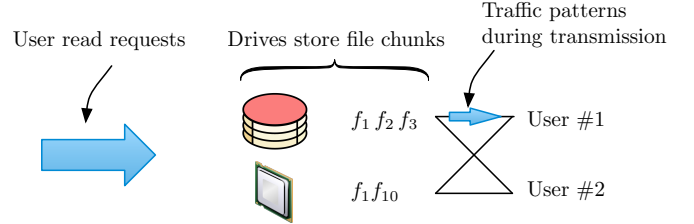


Fig. 1. Illustration of the physical networks we consider in this paper. User requests arrive at the drive network. Drives, of various storage technologies, store some set of file chunks. Time is slotted and during each timeslot, traffic pattern constraints govern how many users each drive can transmit a stored chunk to.

To the author's knowledge, the following key questions are unaddressed when designing and implementing high-traffic storage networks: What is the maximum achievable rate of a storage network with general traffic patterns and arbitrary chunk-to-drive mappings? There do not exist systematic methods of mapping physical networks to queueing models. What is the impact of coded storage on the maximum achievable rate of a storage network? Existing queueing work on coded storage either assumes perfect scheduling or uses scheduling heuristics. Further, what scheduling algorithms achieve maximum rate for coded storage?

Referring to Fig. 2, the main contributions of this paper are as follows.

- We introduce the *queued cross-bar network (QCN)* method, which is a technique to model relatively general physical drive networks as queueing networks. Our technique allows for arbitrary traffic patterns as well as chunk-to-drive mappings.
- To determine the effect of drive traffic pattern restrictions, we develop a framework for analyzing QCN RRs by considering their conflict graph stable set polytopes (SSPs). In doing so, we use and adapt existing techniques from cross-bar switching literature.
- We exactly characterize the SSP of QCN conflict graphs under a variety of traffic patterns, allowing for their efficient RR computation.
- For uncoded storage, we characterize RRs, and prove that the existing scheduling algorithm of Tassioulas *et al.* [3] can be modified to be rate optimal in this application. For coded storage, we develop a RR upper bound, for which we provide an intuitive interpretation. We show that the coded storage RR upper bound is achievable in certain coded systems in which drives store sufficient coded information, as well in certain dynamic coding

systems.

- We present numerical illustrations that show potential increases in RR volume from coded storage, averaging 50% across traffic patterns.

This paper builds upon and complements existing work in coded storage. General scheduling for coded storage in point-to-point networks, when users are served sequentially instead of simultaneously, and with particular file layouts are considered in [4], [5]. Server scheduling is also well studied in matched networks such as cross-bar switches. Throughput-optimal schedules are considered for  $N \times M$  point-to-point cross-bar switches using graph theory and techniques such as the Birkhoff-von Neumann theorem [6]–[8]. Switches with multicast and broadcast capabilities with a queueing analysis flavor are considered in [9]. References [10], [11] attempt to map the multicast problem in cross-bar switches to simpler problems such as block-packing games and round-robin based multicast. Chunk scheduling problems in uncoded peer-to-peer networks, as opposed to point-to-multipoints (PMPs), are considered in [12], and for star-based broadcast networks in [13]. This manuscript differs from these works in that we consider general traffic patterns and arbitrary chunk-to-drive mappings. We also consider scheduling for coded storage.

Scheduling for network coded multicast in multihop wireless networks, using a conflict graph approach, is considered in [14]. Reference [15] developed optimal scheduling algorithms for cross-bar switches with network coding in communications. Enhanced conflict graphs, closely related to classical conflict graphs, were developed to allow analysis of cross-bar switches with network coding. The stable sets of these conflict graphs were then characterized exactly by showing that in certain cases they are perfect graphs. Although in general characterizing stable set polytopes is  $NP$ -hard, if the conflict graph is claw-free, then it can be done in polynomial time [15], [16]. We use the fact that certain traffic patterns produce graphs whose

#### A. Summary of Main Results

All results assume drives with deterministic read times.

- 1) In Sec. IV, we develop a QCN model that can be used to map a storage network with arbitrary file layouts across drives into a queueing network, including systems with nonuniform file or chunk replication. In the spirit of [17], Sec. IV-A describes our QCN model as a moded system, whereby valid modes are described as inequalities that capture multipacket reception, drives with multiple service units, and multiple unicast, broadcast, and multicast traffic or communication patterns. See Fig. 3 for an example.
- 2) Sec. V-A describes the construction of a conflict graph from the QCN model, in which we divide our analysis into systems in which drives have infinite or finite I/O access bandwidth.
- 3) Sec. V-B characterizes the stable set polytope for finite I/O bandwidth systems. We show that, for systems under a multicast traffic pattern, associated conflict graphs are not guaranteed to be claw-free. However, systems with a broadcast only as well as broadcast or single

unicast traffic patterns result in perfect and claw-free conflict graphs. In a system with sufficient multipacket reception or with a multiple unicast traffic pattern, the conflict graph is a quasi-line graph. We then use recent results [18] that allows the characterization of stable set polytopes for quasi-line graphs, which are a strict superset of perfect graphs. See Table I for a summary.

- 4) Sec. V-C adopts and adjusts techniques from Tassiulas *et al.* [3] to transform our conflict graph characterizations into an offline scheduling algorithm that is rate optimal for uncoded storage. For coded storage, Sec. VI develops a RR upper bound, for which we provide an intuitive interpretation, or equivalently the RR given particular dynamic coding systems. The upper bound is found by adding links into an equivalent uncoded QCN model, which intuitively depicts coded storage's additional initial scheduling options.

Sec. VII then presents examples and numerical results showing that the RR of coded storage can subsume that of uncoded storage, with increases in volume averaging 50% across traffic patterns.

The remainder of this paper is organized as follows. The general system model and basic notation is described in Sec. II. Preliminaries are detailed in Sec. III. The QCN model construction is detailed in Sec. IV and the characterization of associated conflict graphs is presented in Sec. V. Sec. VI discusses the effect of coded storage, and Sec. VII presents examples and numerical results. Finally, Sec. VIII concludes the paper.

## II. SYSTEM MODEL

We study storage systems with the following system model.

- File layout: Without loss of generality, consider a single chunked file  $\mathcal{F} = \{f_1, \dots, f_T\}$  is stored in drives, and the  $n$ th drive stores a subset of chunks  $\mathcal{F}_n \subseteq \mathcal{F}$ .<sup>1</sup> (We do not consider multisets in which single drives can store multiple chunk replicas.) Let the total number of chunks stored in the system be equal to  $W = \sum_n |\mathcal{F}_n|$ , and  $\mathcal{F} \subseteq \cup_n \mathcal{F}_n$ .
- Drive behavior: Drives have deterministic read and communication pattern of one chunk per timeslot per service unit. (We do not allow preemption or processor sharing between drives.) Drive  $n$  has  $K_n \in \mathbb{N}_+$  service units vis-à-vis queueing theory. We refer to each service unit as a *virtual drive*, and label the set of virtual drives as  $\{D_1, \dots, D_k, \dots, D_R\}$ , where  $R = \sum_n K_n$ .
- User management: At any given time, the system can manage up to finite  $N$  active users, denoted by  $\mathcal{U} = \{u_1, \dots, u_N\}$ . These  $N$  users can be, for instance, subscribers to a system or connected routers or other aggregating nodes in a larger content distribution network.
- Server behavior: As in classic point-to-multipoint (PMP) networks [19], we consider servers that can multicast chunks read from drives to user subsets with various structures, including a multicast traffic pattern.

<sup>1</sup> $\mathcal{F}$  may represent one or more logical physical files.

Consider a queueing network composed of a set of input queues or buffers  $Q^I$ , each of potentially infinite size, and a set of output lines or sinks  $Q^O$  connecting to outside users. Outside users send read request for file chunks to the network, and requests arrive at  $Q^I$ . When a read request is serviced, appropriate chunks are read from one or more drives, and that read data is then transmitted to a set of users using output lines in  $Q^O$ . We say that a read request has been *served* when that request has left its input queue, the requested chunk has been read from drives and then completed transmission on all appropriate output lines.

All lines have the same capacity called the *line rate*. All virtual drives have the same deterministic capacity and read-times. Time is slotted, where the length of a timeslot is the reciprocal of the line rate plus the read time of a service unit.

### III. PRELIMINARIES

This section will introduce select topics in queueing and graph theory used later in the paper. It will also introduce the reader to the communications or traffic patterns that are considered throughout the paper. Again, refer to Fig. 2 for an illustration of the method used in this paper. Readers fluent in both queueing and graph theory are encouraged to immediately read Sec. IV and to use this section simply as a reference for notation.

#### A. Queueing Theory

This subsection lists preliminary queueing theory definitions used throughout the paper. The reader is referred to [20] for a more thorough survey on queueing theory.

**Definition** A *flow* and *rate* are the stream of all read request chunks, and the average number per timeslot, respectively, that arrive at some input queue  $q \in Q^I$  that need to be serviced via output lines  $Q^O$ . Let  $\mathbf{r} \in \mathbb{R}_+^{|Q^I|}$  denote the *rate vector* of all rates into all input queues.

**Definition** A set of flows is called *admissible* if the sum of the rates of all the flows through each input queue or output line does not exceed one, so inputs and outputs are not oversubscribed.

**Definition** A rate vector is said to be *achievable* if there exists a schedule that can serve it, while keeping all queues stable.

**Definition** The *rate region* is the set of all achievable rate vectors.

#### B. Traffic Patterns

We explore various storage, communication, link, and traffic patterns throughout the paper. In point-to-multipoint (PMP) networks, a number of communication strategies or *traffic patterns* are possible between the physical drives that read chunks, and the users receiving those chunks, depending on the system technologies. Unless explicitly stated otherwise, all systems are assumed to have single packet reception.<sup>2</sup>

<sup>2</sup>A packet will consist of the read chunk, as well as some overhead as required by the communication protocol. We assume that such overheads are negligible.

A PMP storage system has a particular *traffic pattern* if, in every timeslot, the set of feasible flows from virtual drives to output lines must meet particular constraints. General traffic patterns are captured by the queued cross-bar network (QCN) model described in Sec. IV. Specific traffic patterns analyzed in this paper are:

**Definition (Single Unicast)** A system uses a *single unicast* traffic pattern if, in every timeslot, a maximum of one output line can be used to service a chunk read by a single virtual drive.

**Definition (Multiple Unicast)** A system uses a *multiple unicast* traffic pattern if, in every timeslot, each output line can be used to service a chunk, but chunks transmitted along each output line must be read by distinct virtual drives.

**Definition (Broadcast)** A system uses a *broadcast* traffic pattern if, in every timeslot, each output line must transmit the same chunk read by the same virtual drive.

**Definition (Multicast)** A system uses a *multicast* traffic pattern if, in every timeslot, each output line can be used to service a single chunk from any virtual drive.

**Definition (Multipacket Reception)** A system uses  $R_x(j)$ -*chunk multipacket reception* if, in each timeslot, the  $j$ th output line or channel can be used to transmit  $R_x(j)$  unique chunks without error.

From a modeling perspective, multipacket reception (MPR) for communications can be viewed as a generalization of speed-up in cross-bar switches [15]. Speed-up is obtained by the communication medium operating at a faster rate, and multipacket reception can be obtained either by speed-up or by particular communication receivers or codes being employed.

#### C. Coded Storage

Coded storage allows physical drives to store linear combinations of chunks, as opposed to only individual chunk subsets. We refer the reader to [1] for a survey of codes for distributed storage. We denote the  $i$ th uncoded chunk as  $f_i$ , and the corresponding  $i$ th coded chunk in the coded system, stored in the same physical location, as  $f_i^c$ . The set of chunks whose information is encoded or mixed with  $f_i^c$  is called the generation of chunk  $f_i^c$ , whose set of uncoded chunk indices we denote by  $g(i)$ . Chunks, be they coded or not, consume the same storage space (ignoring any overhead for storing coding parameters). In this manuscript, we focus on  $(\alpha, s)$  maximum distance separable (MDS) codes, and do not allow the coded chunks to update or regenerate once they are loaded onto physical drives. MDS codes are those in which a set of chunks or a generation is encoded into  $\alpha$  coded chunks and, if a user downloads any  $s$  coded chunks, then that full generation can be decoded. We make two key assumptions regarding coding:

- No user should receive a replica coded chunk from a generation prior to being able to decode that generation.
- If a chunk  $f_i$  is coded within any generation, then all replicas of  $f_i$  are also coded.

A Reed-Solomon code is an example of an MDS code. During a particular timeslot, a user is said to require  $r$  additional

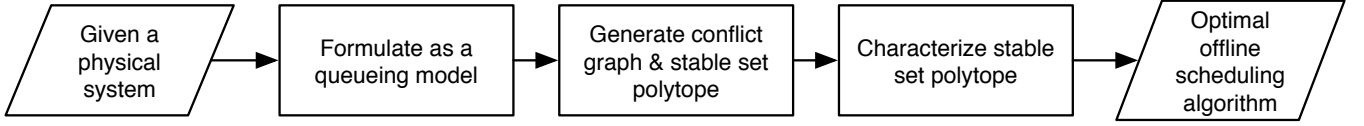


Fig. 2. Summary of the technique used to generate scheduling algorithms in this paper. First, we formulate a physical system as a modeled queueing model. Second, we generate a conflict graph that depicts the constraints on the modes of this queueing model. Third, we apply recent graph theory results that characterize the stable set polytope for this conflict graph. Fourth, we apply known techniques to translate a stable set polytope into an optimal offline scheduling algorithm, in which case incoming traffic statistics are known. The model presented in this paper can also be applied to online scheduling systems in which case no knowledge of incoming traffic statistics is known, by applying policies such as that shown in [3].

*degrees of freedom* if they require  $r$  unique additional coded chunks from the storage system to decode a particular generation to be able to decode chunks in the associated generation. Further,  $f_i^{wc}$  is said to be *innovative* for a user if receiving  $f_i^c$  would reduce the required degrees of freedom for that user by one.

#### D. Graph Theory

This subsection lists graph theory definitions used throughout the paper. We refer the reader to [16] for a more thorough survey on graph theory. Definitions are across graph  $G = (V, E)$ , composed of vertices  $V$  and edges  $E$ .

**Definition (Hyperedge)** A *hyperedge* of graph  $G$  is an edge  $e \in E \subseteq \mathcal{P}(V)$ , the power set of  $V$ . In particular, a hyperedge can connect any number of vertices from that graph, instead of only two vertices.

**Definition (Incidence vector)** The *incidence vector* of a set of vertices  $V_1 \subseteq V(G)$  is a  $\{0, 1\}$ -vector  $\mathbf{x}$  whose entries are labeled with the vertices of  $G$ . If  $x_i = 1$ , then vertex  $i$  is in  $V_1$ ; otherwise,  $i \notin V_1$ .

**Definition (Clique)** A subgraph is called a *clique* if all vertices in the subgraph are pairwise connected.

**Definition (Stable set)** A set of vertices  $V_1 \subseteq V$  forms a *stable set* if for every pair of vertices in  $V_1$ , there is no edge connecting the two.

**Definition (Stability number)** The *stability number*  $\alpha(G)$  is the maximum cardinality of a stable set of  $G$ .

**Definition (Stable set polytope)** The *stable set polytope*  $STAB(G)$  of a graph  $G = (V, E)$  is the convex hull of the incidence vectors  $\mathbf{x}$  of the stable sets of  $G$ .

**Definition (Claw-free graph)** We say that a conflict graph is *claw-free* if no induced subgraph of  $G$  is a vertex with three pairwise disconnected neighbors.

**Definition (Quasi-line graph)** A graph is a *quasi-line graph* if the closed neighborhood of every vertex can be partitioned into two cliques.

**Definition (Chromatic number)** The *chromatic number* of graph  $G$  is the smallest number of colors needed to color the vertices of  $G$  so that no two adjacent vertices share the same color.

**Definition (Perfect graph)** A graph is *perfect* if the chromatic number of every induced subgraph equals the size of the largest clique of that subgraph.

#### E. Conflict Graphs

Conflict graphs are discussed in detail in [21], [22]. A brief overview follows. Given network graph  $G_{\text{net}} = (V_{\text{net}}, E_{\text{net}})$ , and associated feasibility constraints across  $E_{\text{net}}$ , conflict graphs allow the visualization of those feasibility constraints. In this paper conflict graphs are between hyperedges in the queueing network model. In general, conflict graph construction generates a simple<sup>3</sup> and finite conflict graph  $G = (V, E)$  as follows:

- For every possible hyperedge  $e \in E_{\text{net}}$ , create a set of vertices  $v_{(e,j)}$  in  $V$  so that there is a one-to-one correspondence between all possible states  $j$  of hyperedge  $e$  (excluding the empty set state  $\emptyset$ ) and the vertices  $v_{(e,j)}$ .
- Connect vertices  $v_{(e,j)}$  and  $v_{(e',j')}$  if assigning state  $j$  to  $e$  and state  $j'$  to  $e'$  simultaneously is impossible due to a conflict across feasibility constraints [15].

A stable set from the conflict graph  $STAB(G)$  represents a collection of links that can operate simultaneously without conflict, hence it represents a valid system mode. The stable-set polytope (SSP) can be thought of as the convex combination of all valid modes and through timesharing, any point in the SSP can be set as the system operating point.

General conflict graphs have the potential to have a large number of states and to be computationally intractable. Indeed, for general graphs the problem of solving the maximum stable set problem is known to be *NP-hard*. If the graph has particular structure such as being claw-free, then the maximum stable set problem can be solved in polynomial time. However, more than 20 years after the discovery of a polynomial algorithm for the maximum stable set problem for claw-free graphs, the explicit description or characterization of the SSP for claw-free graphs remains an open problem [18]. Recently, it was proved that if the conflict graph is a quasi-line graph (a strict subset of claw-free graphs), then the SSP can be characterized exactly using the clique-family inequalities presented in [18]. We use this result in this paper. In addition, note that if the conflict graph is perfect (a strict subset of quasi-line graphs), then the SSP can also be exactly characterized using the techniques summarized by Kim *et al.* [15].

#### F. Generating Rate Regions from Conflict Graphs

For networks composed of buffers, each of potentially infinite size, and without multicast capabilities, it is well

<sup>3</sup>A graph is simple if it has no loops or parallel edges.

known that the rate region  $\mathbf{R}$  is given by

$$\mathbf{R} = \left\{ \rho \in \mathbb{R}_{0+}^{NT} : \right. \\ \left. \rho \leq \sum_{m \in \mathcal{M}} \phi_m \xi_m, \text{ for some } \phi_m \geq 0, \sum_m \phi_m = 1 \right\}, \quad (1)$$

where  $\xi_m$  are the maximum stable sets of the conflict graph, and  $\mathbb{R}_{0+}^{NT}$  denotes the non-negative real  $NT$ -vectors [17], where  $NT$  is the total number of input buffers in  $\mathcal{Q}^I$ . This is exactly the SSP of the traffic pattern's conflict graph. In addition, it has been shown in [23] that in infinite buffer networks with multicast traffic patterns but no fanout-splitting, the RR is again the SSP of the traffic pattern's conflict graph.

#### G. Generating Scheduling Algorithms from Conflict Graphs

We consider the development of offline scheduling algorithms, which require knowledge of the incoming traffic statistics of read requests into  $\mathcal{Q}^I$ . Given a cross-bar switch network, the stable set polytope from its conflict graph, and a particular operating point that is within the stable set polytope, it has been shown that *frame-based* algorithms, with parameters appropriately chosen, can serve any traffic pattern in the stable set polytope and achieve maximum throughput [15]. A frame is a set of  $F$  consecutive timeslots, where  $F$  is the frame size. Frame-based schedules are specified by a sequence of  $F$  mode schedules and the scheduler cycles through these modes periodically. The authors in [15] generated conflict graphs such that each system queue can be served by a maximum of one vertex, and although this is an assumption we shall generalize, the ideas presented herein will rely heavily on the frame-based offline scheduling from [15].

#### IV. QUEUED CROSS-BAR NETWORK MODEL

This section presents our general queued cross-bar network (QCN) model, which is a queueing model constructed from a physical storage network, as per Sec. II. We generate this queueing network as follows.

- For chunk  $f_i$  and user  $u_j$ , there exists one infinite size input queue labeled  $q_{f_i, u_j}$ . Consider  $\mathcal{Q}^I$  the set of  $T \times N$  input queues.
- For every user  $u_j$  we create an output sink  $q_{u_j}$ , and denote  $\mathcal{Q}^O$  the set of  $N$  output sinks.
- As a reminder of the system model, for each service unit on each physical drive, we create a *virtual drive*  $D_k$ . In the context of the QCN model, when we refer to drives we are always referring to virtual drives.
- For any  $f_i$  that drive  $D_k$  can read, we draw an edge from every queue in  $\{q_{f_i, u_j} \in \mathcal{Q}^I : j \in \{1, \dots, N\}\}$ , i.e., corresponding to chunk  $f_i$  for any user, to output line  $q_{u_j} \in \mathcal{Q}^O$  with label  $D_k$  (labels are not necessarily unique).

System connectivity is represented using three matrices. Let chunk-drive connectivity be defined by  $\mathcal{N}$ , a  $T \times R$  matrix such that each element is defined as

$$\mathcal{N}(i, k) = \begin{cases} 1 & \text{if } f_i \in \mathcal{F}_k \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

During each timeslot, set the status of each user's file knowledge through a  $T \times N \times R$  matrix  $\mathcal{S}$ , such that for a particular timeslot, each element is defined as

$$\mathcal{S}(i, j, k) = \begin{cases} 1 & \text{if chunk } f_i, \text{ that is stored on drive } D_k, \text{ would be} \\ & \text{innovative for user } u_j \text{ in this timeslot} \\ 0 & \text{otherwise.} \end{cases}$$

The *mode set* is given as follows. Let  $\mathcal{M} = \{\mathcal{M}_m\}$  be the set of all modes, where  $\mathcal{M}_m$  is the  $m$ th mode which is an  $T \times N \times R$  matrix where each element is defined as

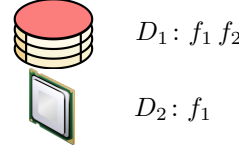
$$\mathcal{M}_m(i, j, k) = \begin{cases} 1 & \text{if user } u_j \text{ receives chunk } f_i \\ & \text{via drive } D_k \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

Let chunk-drive usage indicator  $r_m(i, k)$  be defined as,

$$r_m(i, k) = \begin{cases} 1 & \sum_j \mathcal{M}_m(i, j, k) \geq 1 \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

We call the system the *queued cross-bar network (QCN) model*. See Fig. 3 for a simple illustration.

Example physical system:



Corresponding QCN model:

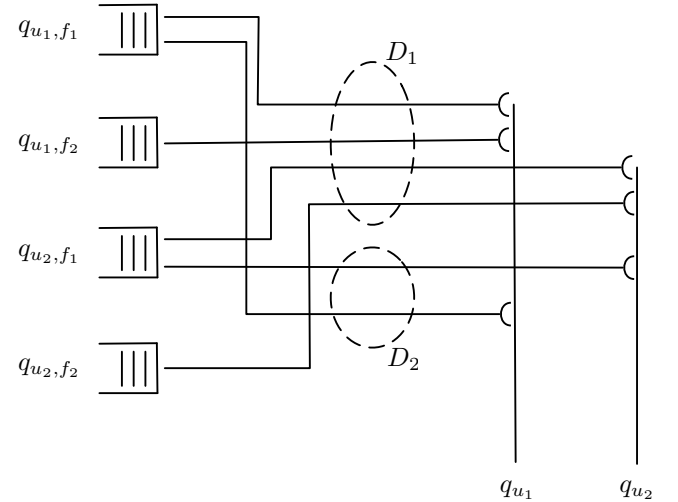


Fig. 3. Example illustration of the queued cross-bar network (QCN) method, modeling a physical system with two users, and two drives storing overlapping chunk sets.

#### A. Constraints

The set of constraints required for mode  $\mathcal{M}_m$  to be *valid* are as follows.

- A user never receives a non-innovative chunk

$$\mathcal{M}_m(i, j, k) \leq \mathcal{S}(i, j, k), \quad \forall i, j, k \quad (5)$$

- Up to  $R_x(j)$  chunks can be received by the  $j$ th user in each timeslot, i.e., user  $j$  has  $R_x$ -chunk multipacket reception capability

$$\sum_{i,k} \mathcal{M}_m(i, j, k) \mathcal{N}(i, k) \leq R_x(j), \quad \forall j \in \{1, \dots, N\} \quad (6)$$

- The  $k$ th (virtual) drive allows up to one read per timeslot

$$\sum_i r_m(i, k) \leq 1, \quad \forall k \in \{1, \dots, R\}. \quad (7)$$

- Traffic pattern constraints:

- Single unicast constraint: Only a single chunk can be transmitted to a single output line

$$\sum_{i,j,k} \mathcal{M}_m(i, j, k) \mathcal{N}(i, k) \leq 1 \quad (8)$$

- Multiple unicast constraint: Only up to one user can receive a chunk  $f_i$  from the same drive  $D_k$

$$\sum_j \mathcal{M}_m(i, j, k) \mathcal{N}(i, k) \leq 1 \quad \forall (i, k) \in \{1, \dots, T\} \times \{1, \dots, R\} \quad (9)$$

- Broadcast constraint: If a read chunk  $f_i$  is transmitted from  $D_k$  to a user, then it is transmitted to all users

$$\exists \mathcal{Y} \subseteq \{(i, k) : \mathcal{N}(i, k) = 1\}$$

s.t.

$$\sum_j \mathcal{M}_m(i, j, k) \mathcal{N}(i, k) = N \quad \forall (i, k) \in \mathcal{Y} \quad (10)$$

$$\sum_j \mathcal{M}_m(i', j, k') \mathcal{N}(i', k') = 0 \quad \forall (i', k') \notin \mathcal{Y}$$

- Multicast constraint: Up to  $N$  users can receive the same chunk from the same drive

$$\sum_j \mathcal{M}_m(i, j, k) \mathcal{N}(i, k) \leq N \quad \forall (i, k) \in \{1, \dots, T\} \times \{1, \dots, R\}. \quad (11)$$

The QCN model incorporates arbitrary chunk layouts, service unit numbers, as well as traffic patterns, which may make it a useful tool in high-traffic storage network analysis. Example systems with the traffic patterns described above are as follows. Some systems may be restricted to multiple unicast traffic patterns if they do not allow caching, others may be restricted to broadcast traffic patterns if they are transmitting wirelessly. Systems can be modeled as having multipacket reception if their scheduling is done at the level of frames, or multiple timeslots if output buffers exist on each output line.

## B. Properties

To explore the storage system types that the QCN model can capture and model, we introduce the following properties.

**Definition (Conservation of flow:)** We say that a queueing model has the *conservation of flow* if in any timeslot, the sum of the number of read requests serviced across input queues  $\mathcal{Q}^I$  is equal to the sum of the number of chunks received at output lines  $\mathcal{Q}^O$ , regardless of traffic pattern constraints.

**Definition (Multiple service unit property:)** We say that a queueing model has the *multiple service unit property*, if for any physical drive  $n$ , up to any fixed  $K_n \in \mathbb{N}_+$  unique and stored chunks can be read from physical drive  $n$  in any timeslot.

A queueing model that has both conservation of flow and the multiple service unit property can be used to model a large variety of systems, regardless of chunk-to-drive layouts.

**Theorem 1.** Any QCN model has conservation of flow.

*Proof:* Any queueing network with fixed topology and in which all service units have deterministic service time has conservation of flow by construction. For a QCN model, in each timeslot a single mode is selected and so we need to check that no mode exists which does not have conservation of flow.

Suppose mode  $\mathcal{M}_m$  does not have conservation of flow. In this case, either we service more read requests from  $\mathcal{Q}^I$  than are received at  $\mathcal{Q}^O$ , or we transmit more chunks down output lines than are serviced across  $\mathcal{Q}^I$ .

Suppose mode  $\mathcal{M}_m$  services more read requests than are received by users. This implies there exists output line  $q_{u_j}$  that receives less chunks than are serviced at queues  $\{q_{f_i, u_j}\}_{i=1}^T$ .  $\mathcal{N}(i, k)$  implies any valid mode can only activate edges on virtual drives with chunks in demand, and owing to (7), there is either overflow servicing from the same chunk on different virtual drives, or from different chunks from different drives. Suppose it is the same chunk on different virtual drives; if those cross bars are activated then  $\sum_k \mathcal{M}_m(i, j, k)$  cross bars are activated. The only way for such invalid overflow to occur is if  $\sum_k \mathcal{M}_m(i, j, k) > R_x(j)$ , and by (6) we have a contradiction. The same contradiction holds for different chunks on different virtual drives.

Suppose mode  $\mathcal{M}_m$  services fewer read requests than are received by users. This implies there exists a queue  $q_{f_i, u_j}$  that transmits a request to more than  $q_{u_j}$ . However, by construction no such labeled links exist in the QCN model, so we have a contradiction. ■

**Theorem 2.** The QCN model has the multiple service unit property.

*Proof:* This follows naturally from the construction of the QCN queueing model from a physical network. If a physical drive  $n$  has  $K_n$  service units, we construct  $K_n$  virtual drives, all with access to physical drive  $n$ 's chunks. All constraints are indexed across virtual drives in  $k$ , so the property holds by construction. ■

## V. RATE REGION CHARACTERIZATION

In this section we generate our conflict graph and then characterize the rate region.

### A. Conflict Graph

Conflict graphs are constructed as follows. We use hyperedges as defined by traffic pattern constraints, so that all data is transmitted along the same fingers of the same hyperedge.

In our conflict graph analysis, we restrict ourselves to edge-based conflict graphs. A conflict graph can be *edge-based* if  $R_x(j) \in \{1, T, T+1, T+2, \dots\} \forall j$ ; for  $R_x(j)$ -multipacket reception with  $1 < R_x(j) < T$ , then the conflict graph may require hyperedges to capture the selection of different copies of chunks.

Drive I/O access bandwidth—tightly coupled to the number of service units per drive—is a key parameter of modern storage systems. As such, to build up intuition, we begin by analyzing simpler systems with infinite I/O access bandwidth, and then move to finite bandwidth systems. An infinite I/O system model would be helpful when individual drive blocking is *not* a bottleneck, and instead traffic patterns are the main constraint so wish to analyze their effects directly.

1) *Infinite I/O Access Bandwidth*: Consider the case when  $K_n \geq T, \forall n$ . In this scenario, a drive can read out all chunks simultaneously in a timeslot, so no intra-drive conflicts can arise.

As defined prior,  $\mathcal{U} = \{u_1, \dots, u_N\}$  is the set of active users in the system. To simplify the problem and reduce vertex numbers, we define valid conflict graph vertices using the set of valid traffic pattern constraints. The connectivity between these vertices is then set by storage and link constraints.

*Vertices*: For every chunk stored in the system  $f_i$ , we ignore the drive that stores it since each drive has infinite I/O bandwidth. Given a particular conflict graph traffic pattern constraint, we generate a set of vertices in our conflict graph as

- Multicast:  $\{v_{f_i, u_S}\}_{S \in \mathcal{P}_{\geq 1}(\mathcal{U})}$ , where  $\mathcal{P}_{\geq 1}(\mathcal{U})$  is the powerset of all active users excluding the empty set of users. The total number of vertices in the conflict graph then scales as  $\mathcal{O}(T2^N)$ .
- Broadcast:  $\{v_{f_i, u_S}\}_{S \in \mathcal{P}_{\geq N}(\mathcal{U})}$ , i.e., a single  $\{v_{f_i, u_{\forall}}\}$  vertex. The total number of vertices in the conflict graph scales as  $\mathcal{O}(T)$ .
- Multiple unicast:  $\{v_{f_i, u_j}\}_{j=1}^N$ , i.e., a set of  $N$  vertices. The total number of vertices in the conflict graph scales as  $\mathcal{O}(TN)$ .

*Edges*: Consider two vertices generated by the traffic pattern constraints,  $v_{f_{i_1}, u_{S_1}}$ , and  $v_{f_{i_2}, u_{S_2}}$ . Given some  $k_1$  and  $k_2$  such that  $\mathcal{N}(i_1, k_1) = \mathcal{N}(i_2, k_2) = 1$ , if setting

$$\mathcal{M}_m(i_1, j_1, k_1) = 1 \quad \forall j_1 \in \mathcal{S}_1 \quad (12)$$

and

$$\mathcal{M}_m(i_2, j_2, k_2) = 1 \quad \forall j_2 \in \mathcal{S}_2 \quad (13)$$

violates at least one storage or link constraint as in Sec. IV-A, then connect  $v_{f_{i_1}, u_{S_1}}$ , and  $v_{f_{i_2}, u_{S_2}}$  with an edge.

2) *Finite I/O Access Bandwidth*: Consider the case when  $K_n < T, \forall n$ .

*Vertices*: Similarly to the infinite I/O scenario, we generate viable vertices via our traffic pattern constraints. The primary addition is that we generate separate vertices for duplicate chunks stored on other virtual drives.

Given our chunk-drive connectivity matrix  $\mathcal{N}$ , for every non-zero element of  $\mathcal{N}(i, k)$  we generate a set of vertices in our conflict graph, given by our traffic pattern constraints as:

- Multicast:  $\{v_{f_{i,k}, u_S}\}_{S \in \mathcal{P}_{\geq 1}(\mathcal{U})}$ , where  $\mathcal{P}_{\geq 1}(\mathcal{U})$  is the powerset of all active users excluding the empty set of users. The total number of vertices then scales as  $\mathcal{O}(TR2^N)$ . As a reminder,  $T$  is the number of chunks,  $R$  the number of virtual drives, and  $N$  the number of users.
- Broadcast:  $\{v_{f_{i,k}, u_S}\}_{S \in \mathcal{P}_{\geq N}(\mathcal{U})}$ , i.e., a single  $\{v_{f_{i,k}, u_{\forall}}\}$  vertex. The total number of vertices scales as  $\mathcal{O}(TR)$ .
- Multiple unicast:  $\{v_{f_{i,k}, u_j}\}_{j=1}^N$ , i.e., a set of  $N$  vertices. The total number of vertices scales as  $\mathcal{O}(TRN)$ .

*Edges*: Consider two vertices generated by the traffic pattern constraints,  $v_{f_{i_1, k_1}, u_{S_1}}$ , and  $v_{f_{i_2, k_2}, u_{S_2}}$ . If setting

$$\mathcal{M}_m(i_1, j_1, k_1) = 1 \quad \forall j_1 \in \mathcal{S}_1 \quad (14)$$

and

$$\mathcal{M}_m(i_2, j_2, k_2) = 1 \quad \forall j_2 \in \mathcal{S}_2 \quad (15)$$

violates at least one storage or link constraint from Sec. IV-A, then connect  $v_{f_{i_1, k_1}, u_{S_1}}$ , and  $v_{f_{i_2, k_2}, u_{S_2}}$  with an edge.

As a simple example of a conflict graph, consider a storage system with two users  $u_1$  and  $u_2$  and two chunks  $f_1$  and  $f_2$  stored on two drives as  $D_1 : f_1$  and  $D_2 : f_2$ . The conflict graph of this system under multicast traffic pattern with  $R_x(1) = R_x(2) = 1$  and drives having single service units is shown in Fig. 4(a). In this graph, as can be seen, since each user can receive only up to one chunk at each timeslot, the vertices that correspond to the same user conflict with each other. In addition, since drives have single service units, the vertices that correspond to the same drive conflict with each other. Furthermore, the conflict graph of the above system under multicast traffic pattern with  $R_x(1) = R_x(2) = 2$  is shown in Fig. 4(b). In this graph, since multipacket reception is allowed, receivers are able to receive up to two chunks per timeslot. Therefore, there is no conflict between the vertices of the same user. However, due to drives' single service units, still the vertices that correspond to the same drive conflict with each other.

### B. Characterizing the SSP

As a reminder, we restrict our characterization to edge-based conflict graphs. We provide characterization analysis for the conflict graphs generated in the prior subsection for finite I/O access bandwidth systems; characterizations for infinite I/O systems are extremely similar.

Unless otherwise stated, we do not allow multipacket reception so  $R_x(j) = 1, \forall j$ . In multipacket reception systems we use the multicast traffic pattern. The framework we have setup thus far provides rapid analysis of many storage systems with various traffic patterns. See Table I for scenarios for which



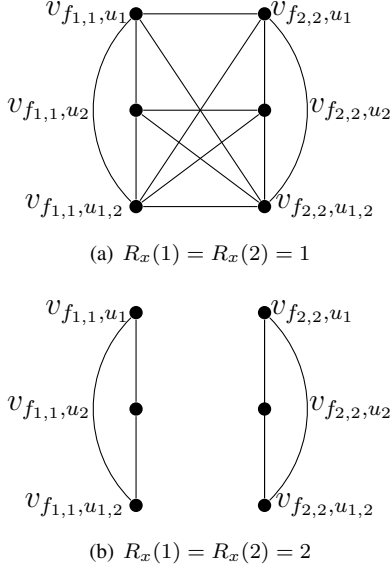


Fig. 4. Conflict graphs of a physical network with two users  $u_1$  and  $u_2$  and two chunks  $f_1$  and  $f_2$ , operating with a multicast traffic pattern.

TABLE I  
QCN MODELS WITH FINITE I/O: TRAFFIC PATTERNS AND THEIR ASSOCIATED CONFLICT GRAPH PROPERTIES.

Traffic Pattern	Claw-free	Notes
Single unicast	Yes	Perfect (Lem. 2)
Broadcast	Yes	Perfect (Lem. 2)
Broadcast and single unicast	Yes	Perfect (Lem. 2)
Multiple unicast	Yes	Quasi-line (Lem. 3)
Multicast	No	(Lem. 1)
Broadcast and multiple unicast	No	(Cor. 1)
$\geq T$ -multipacket reception	Yes	Quasi-line (Thm. 3)
$< T$ -chunk multipacket reception		Unknown

we have characterized the associated conflict graphs, and by extension their associated SSPs.

Given a multicast traffic pattern, the general conflict graph is not claw-free, making exact SSP characterization challenging and motivating the exploration of more restrictive traffic patterns.

**Lemma 1.** *Given a QCN model operating with a multicast traffic pattern, with greater than two users and greater than two chunks, then the associated conflict graph is not guaranteed to be claw-free.*

*Proof:*

Suppose the conflict graph is claw-free. Consider the counter example given by Fig. 5. Given there are  $T \geq 3$  chunks in our target file and  $N \geq 3$  users, consider the broadcast traffic pattern. Owing to the lack of multipacket reception, the set of vertices  $\{v_{f_{i,k},u_{\forall}}\}_{i=1}^T$  form a single clique, where we use the simplifying notation  $u_{\forall} = \{u_1, \dots, u_N\}$ . For a given chunk  $f_{i,k}$ , the set of vertices  $\{v_{f_{i,k},u_j}\}_{i=1}^N$  form a clique, as do vertices  $\{v_{f_{i,k},u_j}\}_{i=1}^T$  for a given user  $u_j$ . In Fig. 5 we depict each of these cliques as a rectangle. As highlighted in red with curved edges, the set of unique vertices

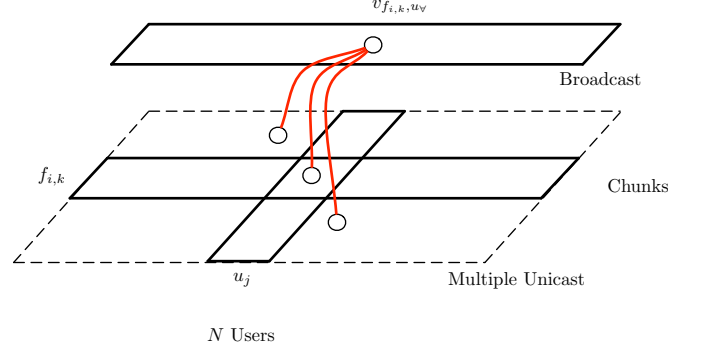


Fig. 5. As per the proof of Lemma 1, given a multicast traffic pattern, an example of a conflict graph with a claw. A clique of vertices is depicted as a rectangle. The graph depicts a subset of the overall conflict graph, illustrating broadcast and multiple unicast clique structure.

$\{v_{f_{i,1},u_{\forall}}, v_{f_{i,1},u_j}, v_{f_{i,2},u_j}, v_{f_{i,3},u_j}\}$  form a claw.

The presented counterexample immediately generates an additional corollary: Systems that allow both multiple unicast and broadcast traffic patterns are not guaranteed to admit claw-free conflict graphs.

**Corollary 1.** *Given a QCN model operating with a traffic pattern that allows either broadcast or multiple unicast, with greater than two users and greater than two chunks, then the associated conflict graph is not guaranteed to be claw-free.*

*Proof:* As per the counterexample shown in Fig. 5.

We now consider systems with restricted traffic patterns, including single unicast, broadcast, and broadcast with single unicast.

**Lemma 2.** *Suppose a QCN model operates with a traffic pattern of either single unicast, broadcast, or broadcast and single unicast. Then its associated conflict graph is claw-free, perfect, and with stability number equal to one.*

*Proof:* First, consider a single unicast traffic pattern. The set of associated conflict graph vertices  $\{v_{f_{i,k},u_j}\}_{i,j,k}$  form a single clique. Second, consider a broadcast traffic pattern. The set of associated conflict graph vertices  $\{v_{f_{i,k},u_{\forall}}\}_{i,k}$  again form a single clique. Third, consider the broadcast or single unicast traffic pattern. The associated conflict graph has one clique from the broadcast and another from the single unicast traffic pattern. Owing to there being no multipacket reception, these two cliques are both fully connected. Cliques have stability number of one and are an example of perfect graphs.

**Lemma 3.** *Suppose a QCN model operates with a multiple unicast traffic pattern. Then its associated conflict graph is a quasi-line graph.*

*Proof:* In constructing the conflict graph, there exist two constraints of note. First, each virtual drive can make up to one transmission, as per (7). Second, each user can receive up to one unicast chunk, as per the multiple unicast traffic pattern. Consider conflict graph vertex  $v_{f_{i,k},u_S}$ . It is a member of two cliques: first, across all chunks on drive  $D_k$  and across all



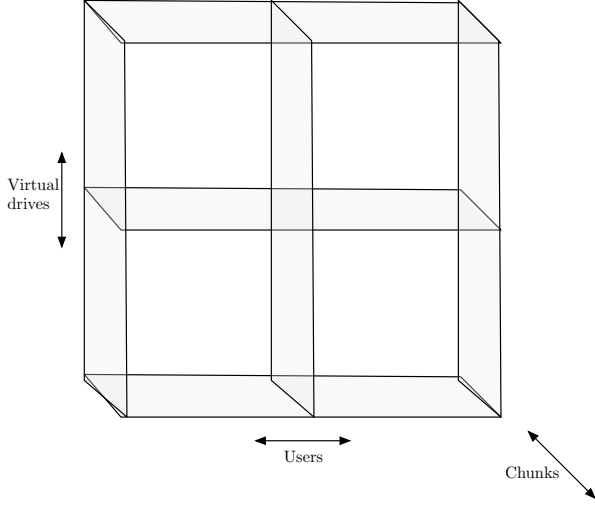


Fig. 6. Simplified visualization of a conflict graph associated with a finite I/O QCN model operating with a multiple unicast traffic pattern. Each horizontal panel represents a clique generated by virtual drive transmissions, and each vertical panel represents a clique generated by user traffic pattern restrictions.

$N$  users. We illustrate this as the horizontal clique panels in Fig. 6. Second, given user set  $u_S$ , all drives and chunks form another clique. We illustrate this as the vertical clique panels in Fig. 6. Not all cliques in Fig. 6 need be of the same size. There are no other connections or conflicts between vertices. The conflict graph is then a quasi-line graph. ■

**Theorem 3.** Suppose a QCN model operates with any traffic pattern, and all users have at least  $T$ -chunk multipacket reception. Then its associated conflict graph is a quasi-line graph.

*Proof:* Consider chunk  $f_{i,k}$ . The vertices generated in the conflict graph associated with  $f_{i,k}$  are a function of the traffic pattern used and the drive service unit constraints. Regardless of traffic pattern particulars, all vertices associated with  $f_{i,k}$  will form a clique owing to drive constraints, i.e., for any  $v_{f_{i,k},u_{S_1}}$  and  $v_{f_{i,k},u_{S_2}}$ , they must be connected, where  $S_1, S_2 \subseteq \{1, \dots, N\}$ . Crucially, however, owing to multipacket reception, unique vertices  $v_{f_{i,k},u_{S_1}}$  and  $v_{f_{j,l},u_{S_2}}$ ,  $i \neq j, k \neq l$  are not connected. The conflict graph is then a set of disjoint cliques, where each clique is associated with a single chunk on a single drive. This set of disjoint cliques is a quasi-line graph as each closed neighborhood can be partitioned into the union of two cliques. ■

We also point out a connection between claw- and *net-free* conflict graphs and QCN models. In graph theory, the study of claw-free graphs is often associated with claw- and *net-free* graphs, which are both claw- and *net-free*. A *net* is illustrated in Fig. 7, which is formed by starting with a triangle and adding to each vertex a new vertex. More is known about claw- and *net-free* graphs than only claw-free graphs. In the generation of our conflict graphs we explicitly do not consider null or do-not-transmit vertices, as per the our hyperedges not including the emptyset. We point out that if do-not-transmit vertices are included in conflict graph generation, then those conflict graphs are not guaranteed to be *net-free*.

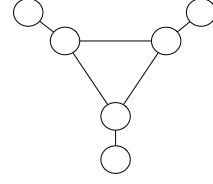


Fig. 7. Net graph illustration. It is formed by starting with a triangle and adding a new vertex to each original vertex.

**Lemma 4.** If do-not-transmit vertices are used in the construction of conflict graph  $G$ , then  $G$  is not guaranteed to be *net-free*.

*Proof:* Consider a QCN model with infinite I/O operating with a broadcast traffic pattern, with three chunks  $T = 3$  on a virtual drive, and one user  $N = 1$ . Using do-not-transmit vertices for the conflict graph generation, we have three transmit vertices  $\{v_{f_{i,1},u_1}\}_{i=1}^3$  which form a central clique. We then also generate three additional vertices  $\{v_{f_{i,1},u_0}\}_{i=1}^3$ , where each vertex in this set denotes not transmitting chunk  $f_{i,1}$ . We then connect vertex pairs  $(v_{f_{i,1},u_1}, v_{f_{i,1},u_0})$  with edges for each chunk, and we have a *net graph*. ■

Given a conflict graph whose SSP can be characterized, as per Table I, we now describe how to characterize the rate regions based on these SSPs.

### C. Characterizing the Rate Region

This subsection establishes the achievable rate region of QCN systems in terms of their associated conflict graphs, where the SSP has been characterized. We find the rate regions by adapting scheduling policy  $\pi_0$  from [3]—which was also adapted to become offline scheduling Algorithm 1 from [15]—to operate on QCN models. The main difference between this paper and policy  $\pi_0$  from [3] is that our approach also allows traffic pattern selection. Further, in Algorithm 1 from [15] each queue in  $\mathcal{Q}^I$  is served by a maximum of one vertex in the conflict graph. In this work, multiple vertices can service any particular queue. As such, we introduce two types of incidence vectors:

**Definition** ( $(f_{i,k}, u_j)$ —incidence vector): An  $(f_{i,k}, u_j)$ —incidence vector  $\mathcal{C}^{i,k,j}$  is a  $\{0, 1\}$ —vector of size  $m$ ,  $m$  being the total number of stable sets in the conflict graph, whose entries are labeled with the stable sets  $S^\ell, \forall \ell \in \{1, \dots, m\}$ . If  $\chi^{S^\ell}(v_{f_{i,k},u_{S \in \mathcal{P}_{\geq 1}(\mathcal{U}_A)}}) = 1$  where  $j \in S$ , then  $\mathcal{C}^{i,k,j}(\ell) = 1$ ; otherwise  $\mathcal{C}^{i,k,j}(\ell) = 0$ .

**Definition** ( $(f_i, u_j)$ —incidence vector): To obtain the  $(f_i, u_j)$ —incidence vector  $\mathcal{C}^{i,j}$ , we add up all  $(f_{i,k}, u_j)$ —incidence vectors  $\forall k \in \{1, \dots, R\}$ ,

$$\mathcal{C}^{i,j} = \sum_{k=1}^R \mathcal{C}^{i,k,j}. \quad (16)$$

Recall that frame-based schedules can be specified by a sequence of  $F$  switch configurations such that the switch cycles through these configurations periodically. These schedules are decided based on prior knowledge of the arrival rates of the

flows, and do not use the instantaneous queue size information to decide the switch configuration.

Our frame-based offline scheduling algorithm is shown in Algorithm 1, which is a modified version of Algorithm 1 in [15] and  $\pi_0$  in [3] using  $(f_i, u_j)$ -incidence vectors.

---

**Algorithm 1** Offline Scheduling Algorithm

---

- 1: Consider a traffic pattern with rate vector  $\mathbf{r}$  and its conflict graph  $G$ . We assume that  $\mathbf{r} \in STAB(G)$ . Then

$$r_{i,j} = \sum_{\ell=1}^m \phi_\ell C^{i,j}(\ell), \quad \forall i, \forall j \quad (17)$$

where  $\sum_\ell \phi_\ell = 1$  and  $\phi_\ell \geq 0, \forall \ell, r_{i,j} \geq 0, \forall i, \forall j$ .

- 2: Assuming all rates  $r_{i,j}F$  and  $\phi_\ell F$  are rational, choose  $F$  such that  $r_{i,j}F$  and  $\phi_\ell F$  are integers for all  $i, j, \ell$ .
  - 3: For each  $\ell$ , use the switch configuration corresponding to  $S^\ell$  for  $\phi_\ell F$  slots. If there are fewer than  $r_{i,j}F$  requests in the queue, then serve all of them. Repeat step 3.
- 

**Theorem 4.** A QCN model that follows Algorithm 1 is stable if and only if the operating point is within the rate region of the QCN model.

*Proof:* To prove that under the offline algorithm the queues  $q_{f_i, u_j}, \forall i, \forall j$  are stable, it is enough to show that the average service rate of queue  $q_{f_i, u_j}$  is always greater than or equal to the arrival rate of flow  $(f_i, u_j)$ . Essentially, (17) expresses the rate  $r_{i,j}$  as a convex combination of the stable sets, which in turn leads to a switch schedule. This is similar to switch schedules generated via the Birkhoff-von Neumann [7] theorem. From (17), it can be seen that on average the summation of the fraction of times allocated to each of the stable sets guarantees that the service time of each queue  $q_{f_i, u_j}, \forall i, \forall j$  is at least equal to the arrival rate of requests to that queue. ■

We note that there exist online scheduling algorithms, which do not require knowledge of incoming traffic statistics, which can be applied to our QCN model. For instance, [3] provides an online scheduling policy that achieves maximum throughput if the arrivals into queues are i.i.d. and independent across incoming flows. Since this property holds in our QCN model, this online policy can be directly applied, in which case the weight assigned to each vertex in the conflict graph is the sum of all queue lengths of the ingress queues to which it is associated.

## VI. THE EFFECT OF CODED STORAGE

Given an uncoded QCN model, we show how to generate an associated coded QCN model, and then compare the rate regions of these two systems. Since analyzing the rate region of coded storage is nontrivial, we develop an upper bound on its rate region instead.

The upper bound for the coded rate region is generated as follows. For each coded chunk  $f_i^c$  stored on  $D_k$  (which is a linear combination of chunks  $\{f_l\}_{l \in g(i)}$  in its generation), from each ingress buffer in the set  $\{q_{f_l, u_j}\}_{j=1}^N$ , add a labeled edge or link to sink  $q_{u_j}$  labeled  $D_k$ . See Fig. 8 as a simple

single user example, where in the uncoded physical system two unique chunks are stored on unique drives. In the coded physical system,  $f_1^c = f_1 + f_2$ , and  $f_2^c = a_1 f_1 + a_2 f_2$ , so we add two edges to the coded QCN model, as compared to the uncoded QCN model. In the general case, in constructing this upper bound we assume that any request in an ingress buffer is successfully serviced by sending any coded chunk that has the requested chunk in its generation, and that no penalty is paid for the user needing to wait to decode the chunk of interest by receiving sufficient degrees of freedom.

The upper bound is equivalent to setting all elements of the user's file knowledge matrix  $\mathcal{S}$  equal to one, in all states. This upper bound can be met if coefficient cycling is performed, where coefficient cycling is the dynamic updating or refreshing of coefficients in a coded chunk such that every degree of freedom request can be served by any coded chunk, as first defined in [24]. Coefficient cycling guarantees reading of chunks from drives without replacement. If chunks are read without replacement within the same generation, then we can then directly apply Algorithm 1. Similar to [25], which considers the achievability of coded storage in systems with an infinite number of storage nodes, we consider achievability given sufficient storage space and chunk layouts on drives. Since it is unclear whether or not coefficient cycling is feasible in all systems, we show that the coded storage upper bound is achievable if all drives store sufficient coded information, even without dynamic coding.<sup>4</sup>

**Lemma 5.** Suppose a QCN model operates with a unicast traffic pattern. The coded storage upper bound is achievable if, for every  $f_i^c$  stored on drive  $D_k$ , then drive  $D_k$  also stores at least  $s - 1$  additional unique coded chunks from the same generation,  $\{f_l^c : l \in g(i), l \neq i\}$ .

*Proof:* The upper bound holds if all ingress buffers can be served by all valid modes across their connected edges, regardless of any user's state via  $\mathcal{S}$ . More specifically, select any ingress buffer  $q_{u_j, f_i} \in \mathcal{Q}^I$  with connected edges labelled  $D_k$ . For any fixed  $(i, j, k)$ , if  $\mathcal{S}(i, j, k) = 1$  in any timeslot, then  $\mathcal{S}(i, j, k) = 1$  must hold in all timeslots. If there exists a read request in  $q_{u_j, f_i}$ , then user  $u_j$  has received less than  $s$  innovative degrees of freedom. Due to the unicast traffic pattern, since  $D_k$  stores at least  $s$  unique coded chunks, then no matter which strict subset of chunks stored on  $D_k$   $u_j$  has already received,  $D_k$  can always serve a coded chunk that is innovative for  $u_j$ . Hence,  $u_j$  can receive an innovative chunk in every timeslot and if  $\mathcal{S}(i, j, k) = 1$  in any timeslot, then  $\mathcal{S}(i, j, k) = 1$  in all timeslots. ■

**Lemma 6.** Suppose a QCN model operates with any traffic pattern. The coded storage upper bound is achievable if, for every  $f_i^c$  stored on drive  $D_k$ , then drive  $D_k$  also stores at least  $s + N$  additional unique coded chunks from the same generation  $\{f_l^c : l \in g(i), l \neq i\}$ .

*Proof:* The upper bound holds if all ingress buffers can be served by all their connected edges and valid modes, regardless

<sup>4</sup>Without loss of generality, we assume each coded chunk is unique and therefore, multisets are not required.

Example physical system:



Corresponding QCN model:

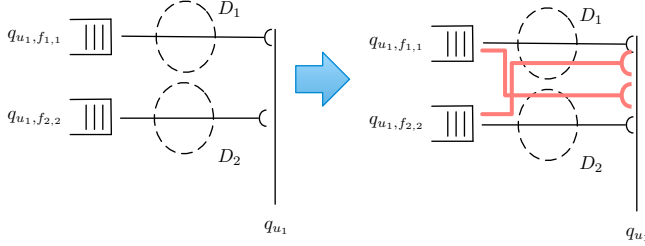


Fig. 8. Example of the intuition behind generating, from an uncoded QCN model, an upper bound for the coded QCN equivalent. In the coded physical network example, two chunks are now coded, and so two additional links are added into the coded QCN model. Intuitively, coding increases the number of links in the QCN system, so a scheduler has more scheduling combinations when routing requests to drives.

of  $\mathcal{S}$ . Consider the ingress queues  $q_{u_j, f_i} \in \mathcal{Q}^I$  with connected edges labeled  $D_k$ . Suppose there exists a subset of users  $N_k \subseteq \{1, \dots, N\}$  for which user state  $\mathcal{S}(i, j, k) = 1, \forall j \in N_k$  in some timeslot, and  $\exists j \in N_k$  such that  $\mathcal{S}(i, j, k) = 0$  in another timeslot. Consider the timeslot in which  $\exists j \in N_k$  such that  $\mathcal{S}(i, j, k) = 0$ . This implies  $D_k$  contains no coded chunk that is innovative for all users  $\{u_j : j \in N_k\}$ . By definition of  $\mathcal{S}$ , all users in  $N_k$  are in a state where they have received less than  $s$  innovative coded chunks. However, the scheduler can then read one of the additional  $N$  coded chunks, since the drive stores  $s+N$  coded chunks from the same generation. We can continue this process for all timeslots until all users have received  $s$  degrees of freedom, so we have a contradiction. ■

Note that the uncoded system rate region is a lower bound for the coded storage. See Sec. VII for examples that compare the uncoded RR with the coded RR upper bound.

## VII. EXAMPLES & NUMERICAL RESULTS

This section walks the reader through the generation of an uncoded system's rate region (RR) under a variety of traffic patterns. All examples are of small storage systems to allow for ease of presentation. We then compare the uncoded RR and coded storage RRs upper bound.

### A. Uncoded RR Examples

This subsection walks the reader through simple RR computation examples using our conflict graph and SSP approach.

**Ex. 1, one chunk, two users under multicast:** Consider a system with two users  $u_1$  and  $u_2$  and a single chunk  $f_1$  stored on drive  $D_1$ , under a multicast traffic pattern. Assuming arrival rate of  $r_{i,j}$  for requests of  $f_i$  from user  $u_j$ ,  $i = 1, j = 1, 2$ , and a general multicast scenario, the conflict graph is shown in Fig. 9. In this conflict graph, three stable sets can be found, where the incidence vectors corresponding to these stable sets

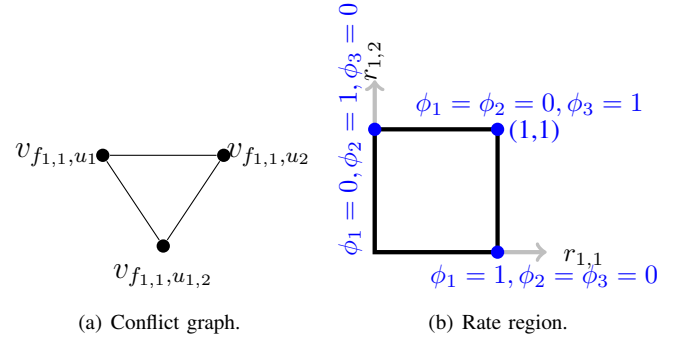


Fig. 9. Conflict graph and RR for Ex. 1, with two users and a single chunk, operating with a multicast traffic pattern.

are as follows

$$\chi^{S_1} = [1, 0, 0], \chi^{S_2} = [0, 1, 0], \chi^{S_3} = [0, 0, 1]. \quad (18)$$

The first, second and third elements of these incidence vectors correspond to vertices  $v_{f1,1,u1}$ ,  $v_{f1,1,u2}$ , and  $v_{f1,1,u1,u2}$ , respectively. Based on the above incidence vectors, the  $(f1, u1)$ - and  $(f1, u2)$ -incidence vectors,  $\mathcal{C}^{1,1,1}$  and  $\mathcal{C}^{1,1,2}$ , can be expressed as

$$\mathcal{C}^{1,1,1} = [1, 0, 1], \mathcal{C}^{1,1,2} = [0, 1, 1]. \quad (19)$$

Since there exists only one drive in the system, the  $(f1, u1)$ - and  $(f1, u2)$ -incidence vectors,  $\mathcal{C}^{1,1}$  and  $\mathcal{C}^{1,2}$ , can be obtained as

$$\mathcal{C}^{1,1} = \mathcal{C}^{1,1,1} = [1, 0, 1], \mathcal{C}^{1,2} = \mathcal{C}^{1,1,2} = [0, 1, 1]. \quad (20)$$

Using (17), the following system of linear equations can be obtained

$$\begin{aligned} \phi_1 + \phi_3 &= r_{1,1}, \phi_2 + \phi_3 = r_{1,2}, \phi_1 + \phi_2 + \phi_3 = 1 \\ r_{1,1} &\geq 0, r_{1,2} \geq 0, \phi_1 \geq 0, \phi_2 \geq 0, \phi_3 \geq 0. \end{aligned} \quad (21)$$

The RR corresponding to this system of linear equations is shown in Fig. 9(b), with area of 1.

**Ex. 2, one chunk, two users under unicast:** Consider Ex. 1 under the unicast traffic pattern, as opposed to multicast. The updated conflict graph is shown in Fig. 10(a). In this conflict graph, two stable sets can be found and the incidence vectors corresponding to these stable sets are as follows

$$\chi^{S_1} = [1, 0], \chi^{S_2} = [0, 1]. \quad (22)$$

The first and second elements of these incidence vectors correspond to vertices  $v_{f1,1,u1}$  and  $v_{f1,1,u2}$ , respectively. Based on the above incidence vectors and the fact that there exists only one drive in the system, the  $(f1, u1)$ - and  $(f1, u2)$ -incidence vectors,  $\mathcal{C}^{1,1}$  and  $\mathcal{C}^{1,2}$ , can be expressed as

$$\mathcal{C}^{1,1} = \mathcal{C}^{1,1,1} = [1, 0], \mathcal{C}^{1,2} = \mathcal{C}^{1,2,1} = [0, 1]. \quad (23)$$

By using (17), the following system of linear equations can be obtained

$$\begin{aligned} \phi_1 &= r_{1,1}, \phi_2 = r_{1,2}, \phi_1 + \phi_2 = 1 \\ r_{1,1} &\geq 0, r_{1,2} \geq 0, \phi_1 \geq 0, \phi_2 \geq 0. \end{aligned} \quad (24)$$

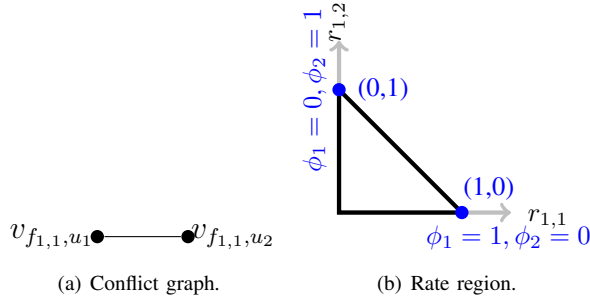


Fig. 10. Conflict graph and RR for Ex. 2, with two users and a single chunk, operating with a unicast traffic pattern.

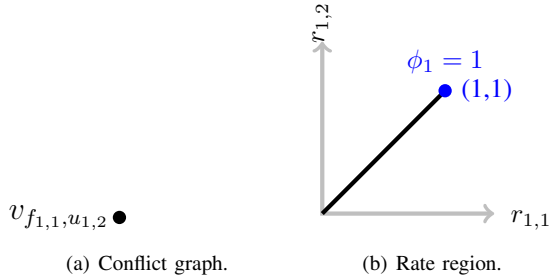


Fig. 11. Conflict graph and RR for Ex. 3, with two users and a single chunk, operating with a broadcast traffic pattern.

The corresponding rate region is shown in Fig. 10(b), with area of  $1/2$ .

**Ex. 3, one chunk, two users under broadcast:** Consider Ex. 1 under the broadcast traffic pattern. This example shows that there are caveats in using RR area/volume as a performance metric that should be carefully considered. The conflict graph is shown in Fig. 11(a), in which only one stable set exists, with incidence vector

$$\chi^{S_1} = [1]. \quad (25)$$

Based on the above incidence vector, the  $(f_1, u_1)$ - and  $(f_1, u_2)$ -incidence vectors,  $\mathcal{C}^{1,1}$  and  $\mathcal{C}^{1,2}$ , are

$$\mathcal{C}^{1,1} = \mathcal{C}^{1,1,1} = [1], \mathcal{C}^{1,2} = \mathcal{C}^{1,1,2} = [1]. \quad (26)$$

By using (17), the following system of linear equations can be obtained

$$\begin{aligned} \phi_1 &= r_{1,1}, \phi_1 = r_{1,2}, \phi_1 = 1 \\ r_{1,1} &\geq 0, r_{1,2} \geq 0, \phi_1 \geq 0. \end{aligned} \quad (27)$$

The corresponding RR is shown in Fig. 11(b), with area of 0.

### B. Comparison of Uncoded and Coded Storage RRs

This subsection compares the RR areas of uncoded storage and the coded storage upper bound. Our coded storage numerical examples use a striped file coded storage layout, which can be seen in [24], and assume  $s/T \in \mathbb{N}_+$ . We first walk the reader through a very simple example that demonstrates the increased scheduling flexibility allowed by coded storage. We then summarize other examples across various traffic patterns.

**Ex. 4, uncoded, two chunks, one user with  $R_x(1) = 2$ :** Consider an uncoded system with one user  $u_1$  and two chunks

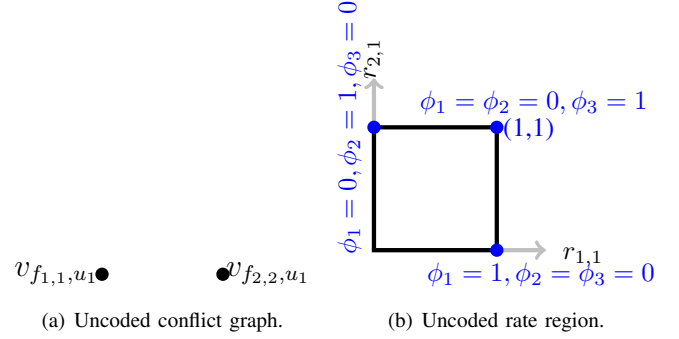


Fig. 12. Conflict graph and RR for uncoded Ex. 4, with one user, two chunks, and multi-packet reception,  $R_x(1) = 2$ .

$f_1$  and  $f_2$  which are stored on drives  $D_1$  and  $D_2$ , respectively. Assuming arrival rate of  $r_{i,j}$  for request of  $f_i$  from user  $u_j$ ,  $i = 1, 2, j = 1$ , and a multicast with multipacket reception setting, the conflict graph is shown in Fig 12(a). In this conflict graph, three stable sets can be found, where the incidence vectors corresponding to these stable sets are as follows

$$\chi^{S_1} = [1, 0], \chi^{S_2} = [0, 1], \chi^{S_3} = [1, 1]. \quad (28)$$

The first and second elements of these incidence vectors correspond to vertices  $v_{f1,1,u1}$  and  $v_{f2,2,u1}$ , respectively. Based on the above incidence vectors, the  $(f_1, u_1)$ - and  $(f_2, u_1)$ -incidence vectors,  $\mathcal{C}^{1,1}$  and  $\mathcal{C}^{2,1}$ , can be expressed as

$$\mathcal{C}^{1,1} = \mathcal{C}^{1,1,1} = [1, 0, 1], \mathcal{C}^{2,1} = \mathcal{C}^{2,1,2} = [0, 1, 1]. \quad (29)$$

By using (17), the following system of linear equations can be obtained

$$\begin{aligned} \phi_1 + \phi_3 &= r_{1,1}, \phi_2 + \phi_3 = r_{2,1}, \phi_1 + \phi_2 + \phi_3 = 1 \\ r_{1,1} &\geq 0, r_{2,1} \geq 0, \phi_1 \geq 0, \phi_2 \geq 0, \phi_3 \geq 0. \end{aligned} \quad (30)$$

The RR corresponding to the above system of linear equations is shown in Fig. 12(b), with area 1.

**Ex. 5, coded, two chunks, one user with  $R_x(1) = 2$ :** Now consider Ex. 4 under a coded storage system, where  $D_1 : f_1 + f_2$  and  $D_2 : a_1 f_1 + a_2 f_2$ . The conflict graph of the coded system is shown in Fig 13(a). In this conflict graph, eight stable sets can be found, where the incidence vectors corresponding to these stable sets are as follows

$$\begin{aligned} \chi^{S_1} &= [1, 0, 0, 0], \chi^{S_2} = [0, 1, 0, 0] \\ \chi^{S_3} &= [0, 0, 1, 0], \chi^{S_4} = [0, 0, 0, 1] \\ \chi^{S_5} &= [1, 1, 0, 0], \chi^{S_6} = [1, 0, 0, 1] \\ \chi^{S_7} &= [0, 1, 1, 0], \chi^{S_8} = [0, 0, 1, 1]. \end{aligned} \quad (31)$$

The first, second, third and fourth elements of these incidence vectors correspond to vertices  $v_{f1,1,u1}$ ,  $v_{f1,2,u1}$ ,  $v_{f2,1,u1}$  and  $v_{f2,2,u1}$ , respectively. Furthermore, based on the above incidence vectors, the  $(f_{1,1}, u_1)$ -,  $(f_{1,2}, u_1)$ -,  $(f_{2,1}, u_1)$ - and  $(f_{2,2}, u_1)$ -incidence vectors,  $\mathcal{C}^{1,1,1}$ ,  $\mathcal{C}^{1,2,1}$ ,  $\mathcal{C}^{2,1,1}$  and  $\mathcal{C}^{2,2,1}$ , can be expressed as

$$\begin{aligned} \mathcal{C}^{1,1,1} &= [1, 0, 0, 0, 1, 1, 0, 0], \mathcal{C}^{1,2,1} = [0, 1, 0, 0, 1, 0, 1, 0] \\ \mathcal{C}^{2,1,1} &= [0, 0, 1, 0, 0, 0, 1, 1], \mathcal{C}^{2,2,1} = [0, 0, 0, 1, 0, 1, 0, 1]. \end{aligned} \quad (32)$$

Therefore, the  $(f_1, u_1)$ - and  $(f_2, u_1)$ -incidence vectors,  $\mathcal{C}^{1,1}$  and  $\mathcal{C}^{2,1}$ , are

$$\begin{aligned}\mathcal{C}^{1,1} &= \mathcal{C}^{1,1,1} + \mathcal{C}^{1,2,1} = [1, 1, 0, 0, 2, 1, 1, 0] \\ \mathcal{C}^{2,1} &= \mathcal{C}^{2,1,1} + \mathcal{C}^{2,2,1} = [0, 0, 1, 1, 0, 1, 1, 2].\end{aligned}\quad (33)$$

Then, by using (17), the following system of linear equations can be obtained

$$\begin{aligned}\phi_1 + \phi_2 + 2\phi_5 + \phi_6 + \phi_7 &= r_{1,1} \\ \phi_3 + \phi_4 + \phi_6 + \phi_7 + 2\phi_8 &= r_{2,1} \\ \phi_1 + \phi_2 + \phi_3 + \phi_4 + \phi_5 + \phi_6 + \phi_7 + \phi_8 &= 1 \\ r_{1,1} \geq 0, r_{2,1} \geq 0, \phi_1 \geq 0, \phi_2 \geq 0, \phi_3 \geq 0, \phi_4 \geq 0, \\ \phi_5 \geq 0, \phi_6 \geq 0, \phi_7 \geq 0, \phi_8 \geq 0.\end{aligned}\quad (34)$$

The corresponding rate region can be obtained as shown in Fig. 13(b), of area 2, whereas the uncoded equivalent has area of 1 (as shown in Fig. 12(b)).

Intuition behind this increase is as follows. Suppose two requests are in queue  $q_{f_1, u_1}$  and none are in  $q_{f_2, u_1}$ . In this case, in the uncoded system servicing this would take two timeslots. In comparison, the coded system could service these two requests in one timeslot.

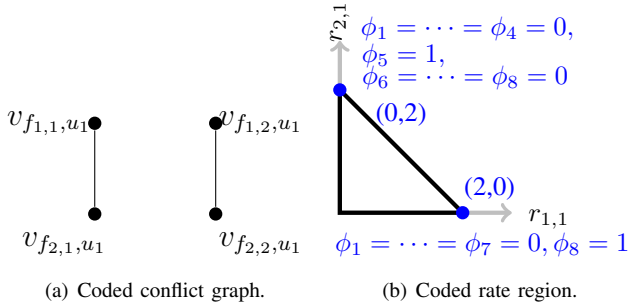


Fig. 13. Conflict graph and RR for coded Ex. 5, with one user, two chunks, and multi-packet reception,  $R_x(1) = 2$ .

**Ex. 6, uncoded and coded, two users, two chunks:** Consider an uncoded system with two users  $u_1, u_2$  and two chunks  $f_1$  and  $f_2$  stored on drives  $D_1$  and  $D_2$ . For the coded system, consider the following drive mapping:  $D_1 : f_1 + f_2$  and  $D_2 : a_1 f_1 + a_2 f_2$ . A comparison of the RR volumes under different traffic patterns and MPR assumptions is presented in Table II. Furthermore, as an example, the conflict graphs for the uncoded and coded systems under a multicast traffic pattern with  $R_x(1) = R_x(2) = 1$  are illustrated in Fig. 14. Note neither graph is a simple clique. For instance, since multicast is allowed, in the uncoded conflict graph, there are no edges between vertices reading from different drives and transmitting to different users, e.g. between  $v_{f_1,1,u_1}$  and  $v_{f_2,2,u_2}$ , and between  $v_{f_1,1,u_2}$  and  $v_{f_2,2,u_1}$ . Similarly, for the coded system, there are no edges between vertices of the form  $v_{f_i,k,u_j}$  and  $v_{f_{i'},k',u_{j'}}$ ,  $i, k, j \in \{1, 2\}$ , where  $k \neq k'$  and  $j \neq j'$ .

**Ex. 7, uncoded and coded, two users, three chunks:** Consider an uncoded system with two users  $u_1, u_2$  and three chunks  $f_1, f_2$ , and  $f_3$  stored on drives  $D_1, D_2$  and  $D_3$ . For the coded system, consider the following drive mapping:

TABLE II  
COMPARISON OF RR VOLUMES FOR EX. 6, 2 CHUNKS, 2 USERS, UNDER UNCODED AND CODED STORAGE.

Traffic Pattern	Uncoded	Coded	% $\Delta$
Single unicast	0.0417	0.0417	0
Multiple unicast	0.1667	0.25	50
Multiple Unicast, with MPR	0.25	0.6667	167
Broadcast	0	0	0
Broadcast, with MPR	0	0	0
Multicast	0.25	0.25	0
Multicast, with MPR	1	2.6667	167
<b>Average</b>			<b>54.8</b>

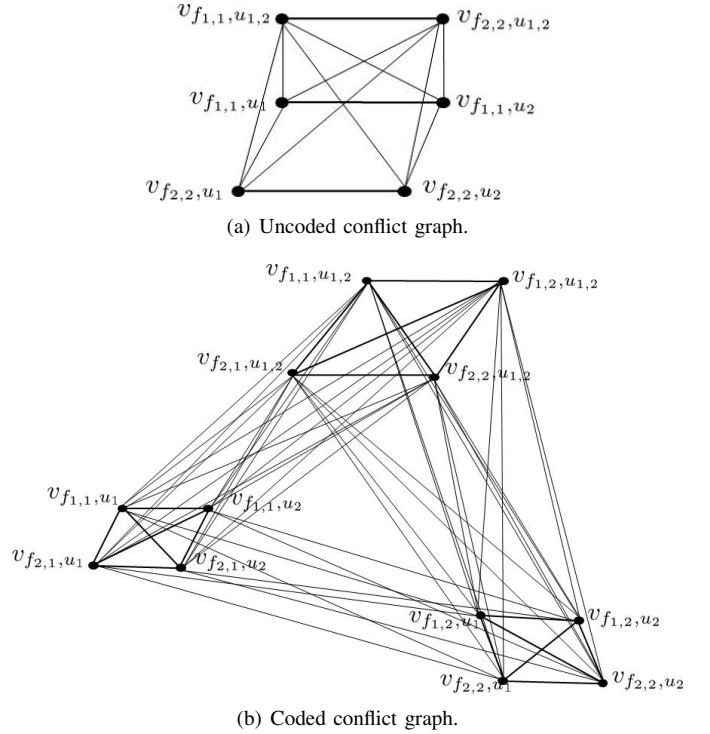


Fig. 14. Conflict graph for uncoded and coded system with two chunks and two users under the multicast setting with  $R_x(1) = R_x(2) = 1$ .

$D_1 : f_1 + f_2 + f_3$ ,  $D_2 : a_1 f_1 + a_2 f_2 + a_3 f_3$  and  $D_3 : a_4 f_1 + a_5 f_2 + a_6 f_3$ , a comparison of the RR's volume for this system under different traffic patterns is presented in Table III.

Results show significant increases in RR volume when using coded storage, averaged across traffic patterns, and as traffic patterns change, the bound shows sizable variability in coded storage gains. This encouraging gain is tempered by the fact that it is an upper bound. Yet, the size of such potential increases warrants further and more exact coded storage RR analysis.

## VIII. DISCUSSION & CONCLUSIONS

Potential areas of future work in this area are as follows. Although the QCN model can be used for arbitrary chunk-to-drive mappings, the seeking within drives is assumed to be deterministic. This may be a reasonable model if drives are of particular solid state varieties, but deterministic drive models for HDDs are problematic. It would be interesting to

TABLE III

COMPARISON OF RR VOLUMES FOR EX. 7, 2 USERS, 3 CHUNKS WITH UNCODED STORAGE AND THE CODED STORAGE UPPER BOUND.

Traffic Pattern	Uncoded	Coded	% $\Delta$
Single unicast	0.0014	0.0014	0
Multiple unicast	0.0236	0.0278	17.8
Multiple unicast, with MPR	0.1250	1.0125	710
Broadcast	0	0	0
Broadcast, with MPR	0	0	0
Multicast	0.0278	0.0278	0
Multicast, with MPR	1	8.1	710
Average			205.4

extend the model to allow for internal finite buffers at drives themselves, as well as arbitrary service distributions.

The conflict graphs generated by the QCN model are a powerful tool for exploring different traffic patterns. However, the state space of the conflict graphs grows quickly in the general case, and it may be useful to explore special cases of storage and traffic patterns that allow for conflict graphs that may not require particular structures that allow for exact characterization, such as quasi-line conflict graphs.

We have developed an upper bound for coded storage rate regions, which is achievable in certain dynamic coding systems, as well as in certain chunk-to-drive mappings. It would be useful to examine the tightness of this bound, as well as to develop a strict lower bound, or to find an exact solution to the rate region of coded storage. Potential avenues include characterizing coded QCN models via user's file knowledge matrix  $\mathcal{S}$  directly.

In conclusion, we have developed a method to map a physical storage system into a simple queued cross-bar network model, with particular application to high-traffic storage systems. In doing so, our method and related analysis tools use existing work in arbitrary queueing networks literature, cross-bar switching, as well as conflict graphs. This allows the QCN method as a natural modeling and analysis tool for systems with non-regular chunk-to-drive mappings, replication, as well as for coded storage. We have used a conflict graph approach, which is a function of storage system traffic patterns, to exactly characterize the stable set polytope of the conflict graphs in a number of cases. We have then computed and compared the rate regions of uncoded storage and of the coded storage upper bound, quantifying promising benefits of coded storage over uncoded systems in terms of RR volume. We have also shown how optimal offline and online scheduling algorithms can be generated from our model.

## REFERENCES

- [1] A. G. Dimakis, K. Ramchandran, Y. Wu, and C. Suh, "A survey on network codes for distributed storage," *Proc. IEEE*, vol. 99, no. 3, pp. 476–489, Mar. 2011.
- [2] D. Ford, F. Labelle, F. I. Popovici, M. Stokely, V.-A. Truong, L. Barroso, C. Grimes, and S. Quinlan, "Availability in globally distributed storage systems," in *Proceedings of the 9th USENIX conference on Operating systems design and implementation*, ser. OSDI'10. Berkeley, CA: USENIX Association, 2010, pp. 1–7. [Online]. Available: <http://dl.acm.org/citation.cfm?id=1924943.1924948>
- [3] L. Tassioulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Trans. Autom. Control*, vol. 37, no. 12, pp. 1936–1948, Dec. 1992.
- [4] N. B. Shah, K. Lee, and K. Ramchandran, "The MDS Queue: Analysing latency performance of codes and redundant requests," *CoRR*, <http://arxiv.org/abs/1211.5405>, 2012.
- [5] L. Huang, S. Pawar, Z. Hao, and K. Ramchandran, "Codes can reduce queueing delay in data centers," in *Proc. IEEE Int. Symp. on Inf. Theory*, Jul. 2012, pp. 2766–2770.
- [6] T. E. Anderson, S. S. Owicki, J. B. Saxe, and C. P. Thacker, "High-speed switch scheduling for local-area networks," *ACM Trans. Comput. Syst.*, vol. 11, no. 4, pp. 319–352, Nov. 1993. [Online]. Available: <http://doi.acm.org/10.1145/161541.161736>
- [7] N. McKeown, A. Mekittikul, V. Anantharam, and J. Walrand, "Achieving 100% throughput in an input-queued switch," *IEEE Trans. Commun.*, vol. 47, no. 8, pp. 1260–1267, Aug 1999.
- [8] C. Caramanis, M. Rosenblum, M. X. Goemans, and V. Tarokh, "Scheduling algorithms for providing flexible, rate-based, quality of service guarantees for packet-switching in banyan networks," in *In Proc. of the 38th annual conf. on info. sciences and systems (CISS)*, 2004, pp. 160–166.
- [9] M. A. Marsan, A. Bianco, P. Giaccone, E. Leonardi, and F. Neri, "Multicast traffic in input-queued switches: optimal scheduling and maximum throughput," *IEEE/ACM Trans. Netw.*, vol. 11, no. 3, pp. 465–477, Jun. 2003. [Online]. Available: <http://dx.doi.org/10.1109/TNET.2003.813048>
- [10] B. Prabhakar, N. McKeown, and R. Ahuja, "Multicast scheduling for input-queued switches," vol. 15, no. 5, pp. 855–866, Jun. 1997.
- [11] H. Yu, S. Ruepp, and M. S. Berger, "Multi-level round-robin multicast scheduling with look-ahead mechanism," in *Proc. IEEE Int. Conf. on Commun.*, Kyoto, Japan, Jun. 2011, pp. 1–5.
- [12] C. Feng and B. Li, *Network coding: Fundamentals and applications*, 1st ed. Academic Press, 2012, ch. Network coding for content distribution and multimedia streaming in peer-to-peer networks.
- [13] G. N. Rouskas and V. Sivaraman, "Packet scheduling in broadcast WDM networks with arbitrary transceiver tuning latencies," *IEEE/ACM Trans. Netw.*, vol. 5, no. 3, pp. 359–370, Jun. 1997.
- [14] D. Traskov, M. Heindlmaier, M. Médard, and R. Koetter, "Scheduling for network-coded multicast," *IEEE/ACM Transactions on Networking (TON)*, vol. 20, no. 5, pp. 1479–1488, 2012.
- [15] M. Kim, J. K. Sundararajan, M. Médard, A. Eryilmaz, and R. Köter, "Network coding in a multicast switch," *IEEE Trans. Inf. Theory*, vol. 57, no. 1, pp. 436–460, 2011.
- [16] D. B. West, *Introduction to graph theory*, 2nd ed. NJ: Prentice Hall, 2001.
- [17] K. Ross and N. Bambos, "Optimizing capacity in interconnection networks with finite buffers," in *Technical Report UCSC-CRL-07-03*, 2007.
- [18] F. Eisenbrand, G. Oriolo, G. Stauffer, and P. Ventura, "The stable set polytope of quasi-line graphs," *Combinatorica*, vol. 28, no. 1, pp. 45–67, 2008.
- [19] K. Foui, J. Casse, I. Sergeev, M. Médard, and M. Maier, "Broadcasting XORs: On the application of network coding in access point-to-multipoint networks," in *MACOM*, 2012.
- [20] L. Kleinrock, *Queueing theory: Theory*. New York: John Wiley & Sons, 1975, vol. 1.
- [21] A. Schrijver, *Combinatorial optimization: Polyhedra and efficiency*. New York: Springer, 2003.
- [22] M. Grötschel, L. Lovász, and A. Schrijver, *Geometric algorithms and combinatorial optimization*. Berlin Heidelberg: Springer-Verlag, 1993.
- [23] J. Sundararajan, S. Deb, and M. Médard, "Extending the Birkhoff-von Neumann switching strategy for multicast - on the use of optical splitting in switches," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 6, pp. 36–50, 2007.
- [24] U. J. Ferner, M. Médard, and E. Soljanin, "Toward sustainable networking: Storage area networks with network coding," in *Proc. Allerton Conf. on Commun., Control and Computing*, Champaign, IL, Oct. 2012.
- [25] S. Acedański, S. Deb, M. Médard, and R. Koetter, "How good is random linear coding based distributed network storage?" in *Proc. 1st Workshop on Network Coding, Theory, and Applications (Netcod'05)*, Apr. 2005.